Confidence belts on bounded parameters

J.Bouchez
DAPNIA/SPP
CEA-Saclay
91191 Gif-sur-Yvette Cedex, France

February 7, 2008

Abstract

We show that the unified method recently proposed by Feldman and Cousins to put confidence intervals on bounded parameters cannot avoid the possibility of getting null results. A modified bayesian approach is also proposed (although not advocated) which ensures no null results and proper coverage.

1 Introduction

In a recent paper [1] , Feldman and Cousins have revisited the long-standing problem of confidence belts on bounded parameters, for which the standard method proposed by Neyman [2] leads in some cases to null, or unphysical, results, and have proposed a new method.

The advocated method takes advantage of the freedom left by the Neyman construction in the choice of ordering used to select the ensemble of values of the measurement with a given probability content.

This new method is strictly classical -or frequentist- (that is, not using bayesian extensions to classical statistics), avoids overcoverage, gives a natural and non-biasing transition between upper limits and intervals, and, according to the authors, avoids null results.

We show in this paper that the last point is not strictly met and that null results can only be avoided for confidence levels above a limit which depends upon the probability law under consideration.

2 The "problem" and its solutions

2.1 Building confidence domains

Let us consider a random variable x of density probability $f(x|\mu)$ where μ is an unknown parameter. Given an observation x, one wishes to make a statement on μ with a given confidence level (noted CL in the following) α. α is the probability that the statement is true.

Let us first consider the case of an upper limit on μ. The Neyman construction consists in defining for each value of μ the value x_m such that

$$F(x_m|\mu) = \int_{-\infty}^{x_m} f(x|\mu)\, dx = 1 - \alpha \tag{1}$$

Whatever μ, x has a probability α to be bigger than $x_m(\mu)$. x being observed, the α CL limit μ_M on μ is obtained by solving $x_m(\mu_M) = x$. In the following, we will consider only the cases where this equation has a unique solution, implying that $x_m(\mu)$ is a monotonic increasing function of μ for any value α. This is equivalent to stating that

$$\frac{\partial F(x|\mu)}{\partial \mu} < 0 \ \vee x, \mu \tag{2}$$

One can equally define an α CL interval $[\mu_m, \mu_M]$ on μ by constructing for each μ an interval $[x_m(\mu), x_M(\mu)]$ in x of probability content α and, given the observation x, solve $x_m(\mu_M) = x_M(\mu_m) = x$. But contrary to upper (or lower) limits, the choice of interval in x is not unique and an ordering prescription on x is necessary.

The usual prescription consists in defining x_m and x_M by

$$\int_{-\infty}^{x_m} f(x|\mu)\, dx = \int_{x_M}^{\infty} f(x|\mu)\, dx = (1-\alpha)/2 \tag{3}$$

(the so-called central interval). This only works for 1-dimensional x, and a new ordering for n-dimensional x has to be chosen. One generally uses

the χ^2 between x and the mean value of x, $\bar{x}(\mu)$, or the likelihood ratio $R = f(x|\mu)/f(x_0|\mu)$ where $f(x_0|\mu)$ is the maximum of f, and one defines a cut c on χ^2 (resp R) so that the probability of $\chi^2 < c$ (resp $R < c$) is α. These two methods (among others), when applied to the 1-dimensional case, would generally give non central intervals.

2.2 The case of bounded parameters

The method outlined above can lead to null results on the parameter μ if this parameter is bounded (for example, $\mu > 0$ if μ is a mass, a variance, etc...). Such a case is illustrated on figure 1, where one sees that for low enough values of x, no upper limit on μ is obtained (or a negative limit is obtained in case $f(x|\mu)$ is defined even for negative μ's).

For a statistician, this is not a problem, since an α CL statement has a $(1-\alpha)$ probability to be wrong. For unbounded parameters, it is impossible to know if a statement is true or false. For bounded parameters, a fraction of the physicists emitting a wrong statement are aware that their statement is wrong, and would rather like not to be in this uncomfortable situation. One should stress however that their result, which when expressed as a limit on μ seems to bring no information, is as legitimate and useful as any other result obtained by those physicists publishing a physical limit (which might be right or wrong).

But the discomfort caused by such a situation is so strong that cures have been searched for to avoid publishing null results.

2.3 The bayesian solution

The bayesian extension to classical statistics consists in building a probability density on the unknown parameter μ from the observation x by applying the (classical) Bayes theorem on conditional probabilities, af if μ were a random variable [3]. This gives a density probability (a degree of belief in bayesian language)

$$g(\mu|x) = K(x)\, f(x|\mu)\, P(\mu) \tag{4}$$

where K is a normalization coefficient insuring that

$$\int_{-\infty}^{\infty} g(\mu|x)\, dx = 1$$

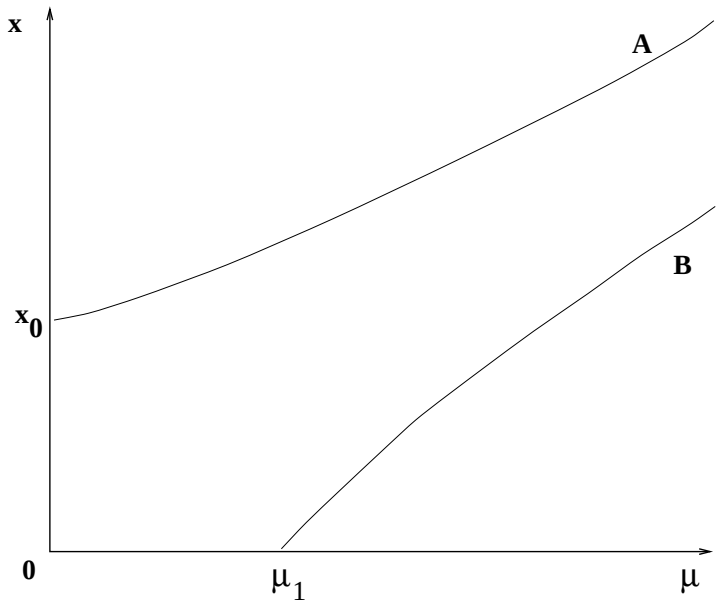

Figure 1: As a fuction of the parameter μ, we show the acceptance domain in x obtained by the frequentist approach (above curve A defined by equation 1) and by the bayesian approach (above curve B defined by equation 5). With curve A, null results are obtained for $x < x_0$, while curve B gives overcoverage.

and $P(\mu)$ summarizes our knowledge on μ prior to the observation x. In particular, for bounded parameters, $P(\mu)$ will be null outside the physical region.

An upper limit at "α CL" μ_0 can then be built on μ by solving

$$\int_{-\infty}^{\mu_0} g(\mu|x)\, dx = \alpha \tag{5}$$

By construction, the selected range of physical μ values will never be empty, so that null results are impossible. This is shown on figure 1.

This approach, although recommended by PDG [4] till 1997, has been criticized by frequentist statisticians for the following reasons:

1. Near the parameter boundary, where null results happen with the standard method, this bayesian "α CL" limit has a classical CL higher than α (it is actually 1 if μ happens to be below μ_1, see figure 1), leading to overcoverage and loss of predictive power. Note however that this overcoverage is a local property, and could well transform into undercoverage at higher μ values.

2. There is some arbitrariness in the choice of $P(\mu)$. PDG proposes to use a Heavyside function to describe the physical limit on μ. Such a prescription has some drawbacks: it is not invariant under changes of parametrization, so that limits will actually depend on the parametrization used. Furthermore, there are cases where the integral in (4) is divergent unless $P(\mu)$ is adequately chosen.

I would like , in the next section, to propose a cure to the second point.

2.4 A modified bayesian limit

A classical upper limit on μ set at μ_0 from an observation x has a confidence level α equal to $1-F(x|\mu_0)$. If one is willing to interpret $P(\mu < \mu_0) = \alpha$ as a probability statement on μ, then the cumulative probability distribution for μ has to be $1 - F(x|\mu)$, so that the probability density for μ deduced from the observation x is given by:

$$\hat{g}(\mu|x) = -\frac{\partial}{\partial\mu} F(x|\mu) \tag{6}$$

Contrary to the usual bayesian definition, $\hat{g}$ is always defined and normalized to 1 by construction. Furthermore, when equation 2 is satisfied, which we have supposed, $\hat{g}$ is positive. When μ is bounded, the definition of conditional probability can be applied to $\hat{g}$ to get its restriction to physical values. If $\mu > a$, then (dropping x in the notation):

$$\tilde{g}(\mu|\mu > a) = \frac{\hat{g}(\mu)}{\int_a^\infty \hat{g}(\mu)\, d\mu} \tag{7}$$

will be the probability distribution of μ restricted to its physical values. $\tilde{g}$ can then be used instead of g (defined in equation 4) to set an upper limit μ_0 on μ, by solving:

$$\int_a^{\mu_0} \tilde{g}(\mu)\, d\mu = \alpha \tag{8}$$

Using $\tilde{g}$ rather than g has several advantages:

- When μ is not bounded, the obtained limit is identical to the classical limit (for bounded parameters, it can be proven it gives overcoverage whatever the value of μ is, contrary to the usual bayesian method).
- the limit is invariant under reparametrization. If we change μ to $\lambda = r(\mu)$, the limit λ_0 will be $r(\mu_0)$
- When prior knowledge is limited to the physical boundary, there is no need to invoke the ambiguous function $P(\mu)$, which is replaced by a Heavyside function irrespective of the parametrization.

Let us note before closing this section that $\hat{g}(\mu|x) = f(x|\mu)$ in some special cases. Among them are the normal law of mean μ and constant variance, and the Poisson law of mean μ, so that for these cases, the limit given by $\tilde{g}$ becomes identical to the former PDG recommendation.

2.5 Feldman and Cousins solution

Feldman and Cousins wanted to fulfill 3 conditions:

1. Avoid null results
2. keep a frequentist approach and produce results with a classical CL equal to α with no overcoverage (except when x is discrete, since this discreteness implies unavoidedly some overcoverage)

3. Solve the "flip-flop" problem of how to switch from upper limit to confidence interval, as they have shown that a choice made according to the value of the observation x leads, as is usually the case, to a biased result, namely an undercoverage (the actual CL being lower than the claimed one). This problem occurs in practice for parameters bounded from below, $\mu \geq a$, where $\mu - a$ is the strength of an hypothetic signal on which to make a statement, and observations also bounded from below ($x > x_0$).

These authors propose an ordering principle on x based on

$$r(x) = \frac{f(x|\mu)}{f(x|\mu_{best})} \tag{9}$$

where μ_{best} is the physical value of μ which maximizes the denominator. $r(x)$ varies between 0 and 1, and one selects for each μ the values of x such that $r(x) > r_c$ and

$$\int_{r>r_c} f(x|\mu)dx = \alpha$$

This construction, being classical, meets the second condition. It will meet the first and third conditions *in cases where* the algorithm selects x values between $x_{min}(\mu)$ and $x_{max}(\mu)$ such that $x_{min} = x_0$ for $\mu < a + s_0$, and $x_{min} > x_0$ for $\mu > a + s_0$, so that an upper limit is published when the observed value of x is below $x_{max}(a)$, and an interval on μ (not necessarily central) for higher values of x.

If it is not the case, the algorithm will lead to null results:

- If x_0 is selected for all values of μ, the algorithm will produce lower limits for $x > x_{max}(a)$ and "null" results below (all physical values of μ being accepted).

- If x_0 is excluded for all values of μ, it will give a null result for $x < x_{min}(a)$, an interval for $x > x_{max}(a)$ and an upper limit inbetween.

We show in the next section that such a situation is likely to occur.

2.6 Study of the ordering algorithm

In order to prove the above mentionned fact, it is sufficient to exhibit cases where null results can be obtained.

I will restrict in the following to cases where μ and x both take only positive values, and where $f(x|\mu)$ is , up to a normalization factor, a function $g(y)$ with $y = x/\mu$. If $\int g(y)dy = 1$, the normalization factor is $1/\mu$, so that $f(x|\mu) = g(y)/\mu$. (Such cases can be met in real cases: y can follow a χ^2 law and μ be the unknown variance, x then being the unnormalized χ^2 from which one wishes to give a statement on the variance). Note that the limits on x at any confidence level are just straight lines passing thru the origin, and no problem of null results is encountered with the usual classical method. The problems appear only when for some reason, one restricts μ to values higher than a, a strictly positive constant (in the χ^2 example, one knows that the variance is at least a, and one wishes to know if there is some other contribution to it). Note also that such distributions satisfy the condition of equation 2.

When no bound is imposed on μ (other than being positive), μ_{best} as defined in (9) is μ_0 given by

$$\frac{\partial}{\partial\mu} f(x|\mu_0) = 0$$

This equation is equivalent to

$$\frac{d}{dy} h(y_0) = 0$$

where $h(y) = yg(y)$ and $y_0 = x/\mu_0$.

Let us suppose that $h(y)$ is such that it has a unique maximum at $y = y_0$.

To build $r(x)$ when $\mu > a$, two cases have to be considered:

- when $x > ay_0$, $\mu_0 > a$, $\mu_{best} = \mu_0$ and $r(x) = h(y)/h(y_0)$
- when $x < ay_0$, $\mu_0 < a$, $\mu_{best} = a$ and $r(x) = h(y)/h(\mu y/a)$

Note that $r(x_0 = \mu y_0) = 1$, and

$$r(0) = \lim_{y\to 0} \frac{h(y)}{h(\mu y/a)} = \left(\frac{a}{\mu}\right)^{(n+1)}$$

where n is the first non-null derivative of $g(y)$ at $y = 0$.

When $\mu \to a$, $r(x) \to 1$ for x between 0 and ay_0, then decreases.

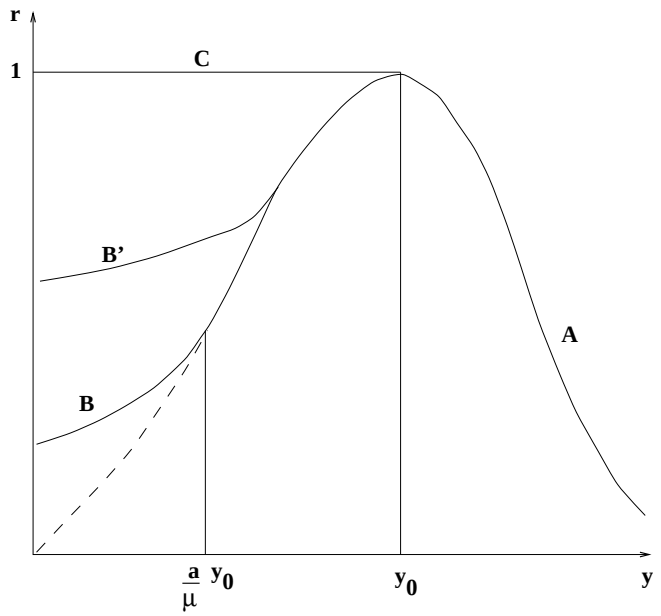

Figure 2: r is shown as a function of y. $h(y)$ takes its maximum at y_0. The function r is made of 2 curves, curve A given by $h(y)/h(y_0)$ when $x > ay_0$, curve B given by $h(y)/h(\mu y/a)$ when $x < ay_0$. The curve B' corresponds to a value μ' of μ closer to a, and curve C is obtained when $\mu = a$.

$r(x)$ is shown on figure 2. We will suppose in the following that $r(x)$ is a non decreasing function of x for x between 0 and ay_0, as it is the case in the explicit examples below (this is to avoid disconnected acceptance domains).

Then, the acceptance domain for x contains $x_0 = \mu y_0$ and extends to both sides according to the chosen value of α, including or not $x = 0$. When $\mu \to a$, the acceptance domain will or not contain $x = 0$ depending on the value of $I = \int_0^{y_0} g(y)dy$. I is ≤ 1. If I is smaller than α, the acceptance domain at $\mu = a$ will include $x = 0$ and no null results are possible. If I is bigger than α, the acceptance domain at $\mu = a$ will NOT include 0 and null results will occur.

One thus sees that the occurence or not of null results with Feldman and Cousins method will depend on the chosen CL. Contrary to the bayesian approach, the absence of null results is not insured by construction.

2.7 Some examples

2.7.1 exponential law

$g(y) = e^{-y}$

$h(y) = ye^{-y}$ has a unique maximum at $y = 1$.

$I = \int_o^1 g(y)dy = 1 - e^{-1} = 0.63$

A CL higher than 0.63 will avoid null results.

This law is a χ^2_2 law (2 degrees of freedom) for $2y$. More generally, χ^2_N laws give values of I slowly decreasing to $1 - 3e^{-2} = 0.594$ when $N \to \infty$.

Other usual laws tend to give values of I smaller than the usual choices of CL, 0.9 or higher, so that in practical cases, null results are likely to be avoided. It is however possible, by a careful choice of $g(y)$, to obtain situations where I can be made as close to 1 as one wishes. For example, the sigmoid-like function:

$$g(y) = \frac{k}{e^{(y-b)} + 1}$$

with y and b positive gives for large values of b an integral I whose value approaches $1 - \log(b)/b$ so that it can be made higher than any given CL.

The next example, purely academic, exhibits a case where null results are unavoidable.

2.7.2 flat distribution

$g(y) = 1$ for y between 0 and 1

x in this case is flat between 0 and an unknown value μ restricted to be bigger than a. $r(x)$ in this case is equal to a/μ for $x \leq a$, and equal to x/μ for x between a and μ. The acceptance domain for x goes from $(1-\alpha)\mu$ to μ (by connexity of the interval, higher values of x being selected first).

Thus, whatever the chosen value for α, $x = 0$ will be excluded from the acceptance domain for all values of μ, only lower limits will be obtained and null results will always occur.

3 Discussion

We have shown in this paper that the newly proposed method to put confidence belts on bounded parameters fails to meet some of its advocated properties, at least in principle although not in practice for most cases.

Why should such a method be preferred to others? (Note that PDG, in its 1998 edition [5], recommends this new method). Feldman and Cousins argue that their method disantangles estimation and hypothesis testing, but one can argue as well that they obtain estimations from an ordering based on a quantity used for hypothesis testing! In view of the failure to avoid null results, I don't think any real argument can be given if one wishes to stick to classical statistics. In my opinion, different methods, as long as they give actual confidence levels equal to the announced one, are mathematically equally acceptable, and a comparison between them implies, consciously or not, some dose of "bayesian" input.

The main problem at hand was to avoid null results. To cure this problem (completely for the bayesian solution, partly only with Feldman and Cousins), one is led to break the symetry of central intervals which equally separates the wrong statements between the victims of fluctuations towards high values of x and victims of fluctuations towards low values. To avoid null results, a dissymetry is introduced to favor the latter to the detriment of the former, who anyway will never know if they are right or wrong. One could argue it is not fair!

I would like to add that the good points of Feldman and Cousins' method could be translated to the bayesian approach to get rid of all the criticisms

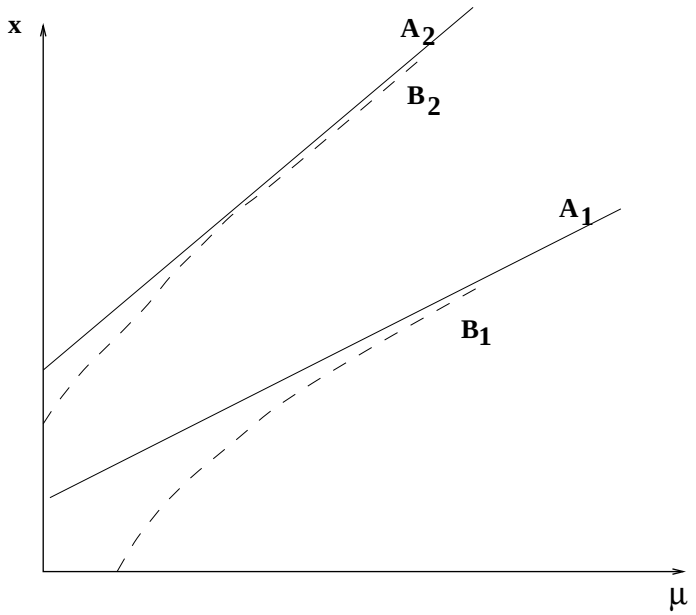

Figure 3: The classical belt at α CL is given by curves A1 and A2 defined by eqution 3. B1 is the bayesian upper limit obtained from formula 8 for a CL equal to $(1+\alpha)/2$. It is complemented by the curve B2 so that for each μ value, the probability content for x between B1 and B2 is α.

it has received. We have already shown in section 2.4 how to solve the criticisms of arbitrariness by using a modified probability density for μ. The only critics left is that of overcoverage for bayesian upper limits. But what Feldman and Cousins have shown is that one should publish intervals, which happen to be upper limits when the interval starts at the physical boundary. Thus, it would be perfectly legitimate, in order to avoid overcoverage, to build the "$(1+\alpha)/2$" upper limit with our modified bayesian recipee; it will show overcoverage, but this can be removed by implementing a lower limit on μ obtained by defining $x_{max}(\mu)$ so that the probability content of selected x be α for any μ. Such a construction would be devoid of any criticism : no overcoverage, no null results, no arbitrariness, and it concides with the classical central intervals of the Neyman construction for unbounded parameters. This is illustrated on figure 3.

The results obtained for Poisson statistics (in the case of a signal over a known background) would be the same as the usual $(1+\alpha)/2$ (instead of α, but this is the price to be paid to avoid flip-flop biasing) bayesian upper limit suitably complemented by a lower limit at high values of x, both curves conciding with the Neyman usual central interval for zero background.

This method I think would certainly satisfy the advocates of the bayesian approach, and be acceptable by classical statisticians. Its construction might look kind of odd, but it uses as well as others the freedom left in the Neyman construction while fully responding to the anxiety created (with no good reason in my opinion) by the occurence of null results.

I will close this paper with a final remark.

All the recent discussions on the choice of method to be used, linked to the fact that different conclusions could be reached (for example, whether or not the recent Karmen result [6] excludes the LSND result [7] on neutrino oscillations [8]) have taken such an importance due to the tendancy to publish results at lower and lower confidence levels. Let me remind that the probability that 2 results being published at 90% CL are both right is only 0.81 ! And with 90 % CL, a non negligible fraction of experiments can obtain null results. It would be in my opinion much more reasonable to publish results with CL of at least 99%, so that the intensity of discussions would considerably drop, because on the one hand, the fraction of null results would reach a very low level and on the other hand the probability of two results being both right would reach 98 %, whatever the algorithms used. Certainly results would look less spectacular, but would gain in fiability.

References

[1] G.J.Feldman and R.D.Cousins, Phys.Rev. D57 (1998) 3873

[2] J.Neyman, Philos.Trans.R.Soc.London A236 (1937) 333

[3] See for example W.T.Eadie et al, statistical methods in experimental physics, North Holland publishing company, 1971, chapter 6

[4] Particle Data Group, R.M.Barnett et al., Phys.Rev. D54 (1996) 162

[5] Particle Data Group, Eur.Phys.J. C3 (1998) 174-177

[6] J.Kleinfeller, 6^{th} international workshop on topics in astroparticle and underground physics (TAUP99) Paris, sept.6-10, 1999 to be published in Nucl.Phys.Proc.Suppl.

[7] C.Athanassopoulos et al., Phys.Rev. C54 (1996) 2685

[8] K.Keitel, hep-ex/9909036